

# Learning Markov Network Structure using Brownian Distance Covariance

Ehsan Khoshgnauz\*

March 8, 2013

## Abstract

In this paper, we present a simple non-parametric method for learning the structure of undirected graphs from data that drawn from an underlying unknown distribution. We propose to use Brownian distance covariance to estimate the conditional independences between the random variables and encodes pairwise Markov graph. This framework can be applied in high-dimensional setting, where the number of parameters much be larger than the sample size.

## 1 Introduction

Undirected graphical models, also known as Markov random fields or Markov networks, have become a part of the mainstream of statistical theory and application in recent years. These models use graphs to represent conditional independences among sets of random variables. In these graphs, the absence of an edge between two vertices means the corresponding random variables are conditionally independent, given the other variables. Learning the structure of a graph is equivalent to learning if there exists an edge between every pair of nodes in the graph.

In the past decade, significant progress has been made on designing efficient algorithms to learn undirected graphs from high-dimensional observational datasets. Most of these methods are based on either the penalized maximum-likelihood estimation or penalized regression methods. Works has focused on the problem of estimating the graph in this high dimensional setting, which becomes feasible if graph is sparse.

---

\*ehsankhnz@gmail.com, Department of Statistics, Tehran Payame Noor University, Tehran, Iran.

[Friedman (2007)] develop an efficient algorithm for computing the estimator with excellent theoretical properties using a graphical version of the lasso.

In high dimensional problems normality assumption is the main constraint in methods. But we can replace the Gaussian constraint with a semi-parametric Gaussian copula, as discussed in [Liu, et al (2009)]. This method use a semi-parametric Gaussian copula or non-paranormal approach by replacing linear functions with a set of one-dimensional smooth functions, for high dimensional inference. The non-paranormal extends the normal by transforming the variables by smooth functions.

The Gaussian distribution is almost always used in study for this scope, because the Gaussian distribution represents at most second-order relationships, it automatically encodes a pairwise Markov graph. The Gaussian distribution has the property that if the  $ij$ th component of inverse covariance matrix is zero, then variables  $i$  and  $j$  are conditionally independent, given the other variables.

Distance correlation is a measure of dependence between random vectors introduced by [Rizzo, and Bakirov (2007)]. For all distributions with finite first moments distance correlation, it is zero if and only if the random vectors are independent. We use this properties to construct our method.

In this paper, we have discussed the structure learning of Markov graphs with large-dimensional covariance matrices where the number of variables is not small compared to the sample size. It is well-known that in such situations the usual estimator, the sample covariance matrix, may not be invertible. The approach suggested is to use distance covariance matrix towards the identity this matrix.

## 2 The proposed method

In this paper we are concerned with the task of estimating the graph structure of a Markov random field over a random vector  $X = (X_1, X_2, \dots, X_p)$ , given  $n$  independent and identically distributed samples.

It was shown that the distance covariance is zero if and only if the two vectors were independent, we use this property.

The main idea is to create a matrix from each pair of distance correlation. Then use it to construct an adjacency matrix of conditional independents between each node pair.

The distance dependence statistics in [Rizzo, and Bakirov (2007)] are defined as follows. For a random sample  $(X, Y) = (X_k, Y_k) : k = 1, \dots, n$  of  $n$  i.i.d. random vectors  $(X, Y)$  from the joint distribution of random vectors  $X$  in  $\mathbb{R}^p$  and  $Y$  in  $\mathbb{R}^q$ , compute the

Euclidean distance matrices  $(a_{kl}) = (|X_k - X_l|_p)$  and  $(b_{kl}) = (|Y_k - Y_l|_q)$ . Define  $A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}$ ,  $k, l = 1, \dots, n$ , where

$$\bar{a}_{k.} = \frac{1}{n} \sum_{l=1}^n a_{kl}, \quad \bar{a}_{.l} = \frac{1}{n} \sum_{k=1}^n a_{kl}, \quad \bar{a}_{..} = \frac{1}{n^2} \sum_{k,l=1}^n a_{kl}. \quad (1)$$

Similarly define  $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$ ,  $k, l = 1, \dots, n$ , Then sample distance correlation  $dcor(X, Y)$  are defined by,

$$R^2(X, Y) = \begin{cases} \frac{\nu^2(X, Y)}{\sqrt{\nu^2(X)\nu^2(Y)}}, & \nu^2(X)\nu^2(Y) > 0; \\ 0, & \nu^2(X)\nu^2(Y) = 0. \end{cases} \quad (2)$$

where

$$\nu_n^2(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl} B_{kl} \quad (3)$$

is the distance covariance.

Also, distance covariance has simple computing formula, the computations would appear to be  $O(n^2)$ , which can be burdensome for large  $n$ .

This estimator is distribution-free and has a simple explicit formula that is easy to compute and interpret.

In our study this formula become more easier, because here  $p = q = 1$ , for example in procedure to form distance correlation matrix,  $a_{kl}, b_{kl}$  for each vector only one time computed, furthermore  $\sum_{k,l=1}^n$  is changed to  $2\sum_{k < l; k,l=1}^n$  and requires less computing time. In Appendix A, we show a function coded in R, that calculate this.

The distance correlation, is implemented in the R package energy [2].

We construct a matrix R of sample distance correlation ( $dcor$ ) between each pair of nodes, so the element  $i, j$  in R is equal to  $dcor(X_i, X_j)$ ,

$$R = \begin{pmatrix} 1 & dcor(X_1, X_2) & \dots & dcor(X_1, X_p) \\ dcor(X_2, X_1) & 1 & \dots & dcor(X_2, X_p) \\ \vdots & \vdots & 1 & \vdots \\ dcor(X_p, X_1) & dcor(X_p, X_2) & \dots & 1 \end{pmatrix} \quad (4)$$

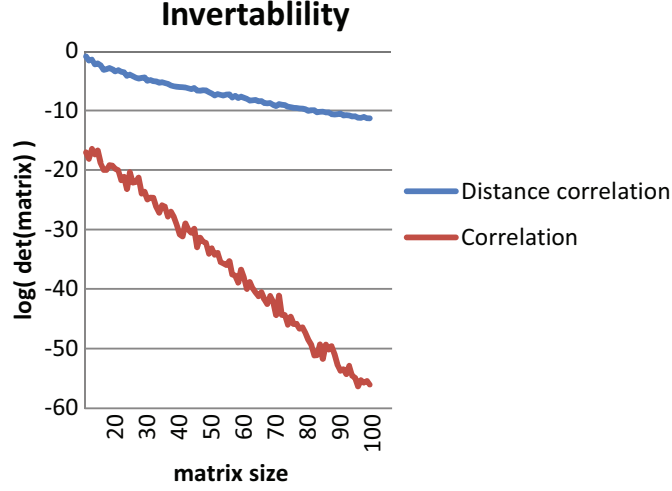


Figure 1: The logarithm of determinant of correlation matrices in different dimensions.

We call the matrix  $R$  defined above the distance correlation matrix.

From the property that  $dcor(Y, X) = dcor(X, Y)$ , obviously  $R$  is a symmetric matrix. When the matrix dimension  $p$  is larger than the number  $n$  of observations available, the ordinary sample covariance matrix is not invertible. But distance covariance or distance correlation matrix, has better performance.

We generate three different random data in 2 to 100 dimensions, and compute average of determinant of the correlation and distance correlation matrix in each dimension. As we see these results in figure 1, distance correlation is more invertible. and computationally, is non-singular enough.

Now we construct The partial correlation matrix. When  $r_{ij}|REST$  be the partial correlation between the variables  $X_i$  and  $X_j$ , given all the remaining variables, and  $P = R^{-1}$ ,  $P = (p_{X_i X_j})$ , the inverse of the correlation matrix [Whittaker (1990)]; is given by,

$$\rho_{X_i X_j | REST} = \frac{-p_{ij}}{[p_{ii} p_{jj}]^{\frac{1}{2}}}. \quad (5)$$

We can calculate matrix  $P$ , by simple computations. As in `partial.cor` function in R Package `Rcmdr`[1] implemented, following code in R language programming do this.

```
RI <- solve(R)           #RI is the inverse of R
D <- 1/sqrt(diag(RI))
P <- -RI * (D %o% D)     #%o% is the outer product operator
diag(P) <- 0
```

Finally  $P$  is the sparse structure matrix of the graph. We can compare each element of  $R$  to a tuning parameter (forming the paths) and derive desired adjacency matrix.

### 3 Simulation Results

In this simulation, we demonstrate the performance of the proposed approach on finding the sparse structures of random Markov networks, by generating Erdős-Rényi random graphs.

The Erdős-Rényi random graph  $G_p$  is a graph on  $p$  nodes in which the probability of an edge being in the graph is  $\frac{c}{p}$  and the edges are generated independently. In this random graph, the average degree of a node is  $c$ .

Given the precision matrix for a zero-mean Gaussian distribution, it is easy to sample data from the distribution. But we do not know the distribution. So we randomly constructed precision matrices, and set random linear relationships with white noise between the columns of data sample matrices.

In a similar manner to [Lin, et al (2009)], We simulated Erdős-Rényi random graphs in two types of sparse structures (or precision matrices): 1) 50 nodes with averagely 3 neighbours per node, 2) 200 nodes with averagely 4 neighbours per node.

Now, the goal here is to see how well our approach recovers the sparse structures of those

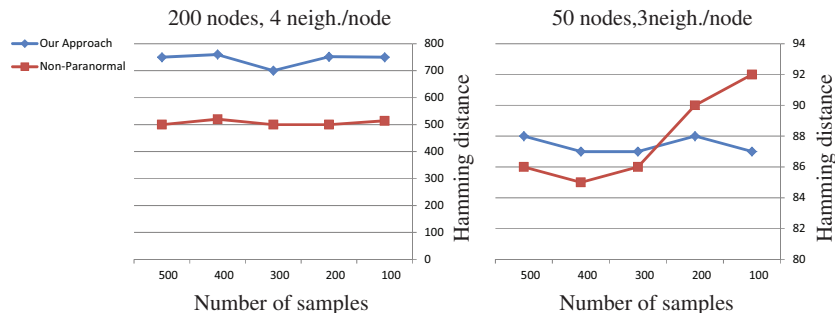


Figure 2: The performance comparison between our proposed approach and the non-paranormal approach.

precision matrices given different numbers of sampled data.

Figure 2 illustrates the performance of our approach in recovering the structures of different Markov networks in comparison with the non-paranormal approach as discussed in [Liu, et al (2009)]. The performance is evaluated by Hamming distance, the number of disagreeing edges between an estimated network and the ground truth, in an equal number of edges.

## 4 Experimental Results

In this section, we are compared our algorithm to that of non-paranormal method as discussed in [Liu, et al (2009)], using the function `huge.npn()` implemented in the R package `huge` [Zhao, et al (2012)] for estimating a semi-parametric Gaussian copula model by truncated normal or normal score.

The example is based on a stock market data which is contributed to the `huge` package that shows closing prices from all stocks in the *S&P 500* for all days that the market was open between January 1, 2003 and January 1, 2008. This gave us 1258 samples for the 452 stocks that remained in the *S&P 500* during the entire time period. Here for convenience and more visibility, we select only first 20 parameters. Also and instead of force-based graph drawing layout of Fruchterman-Reingold that utilized in plot function in `huge`, we use R package `ggm`[3] for visualize graphs more distinctly.

The output of `huge` package graph estimation using the transformed data method and also preprocessing step that mentioned in [Zhao, et al (2012)] is shown in Figure 3.

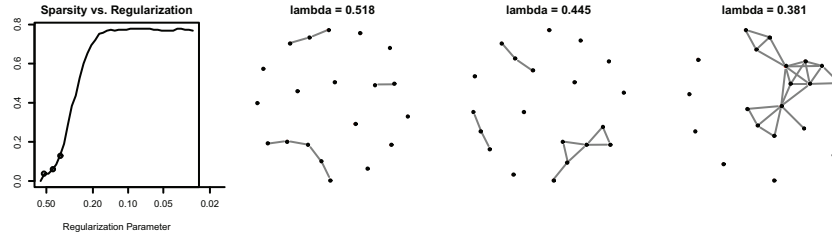


Figure 3: The output of huge, when  $n\lambda = 40$ ,  $\lambda_{\min} \text{ratio} = .05$

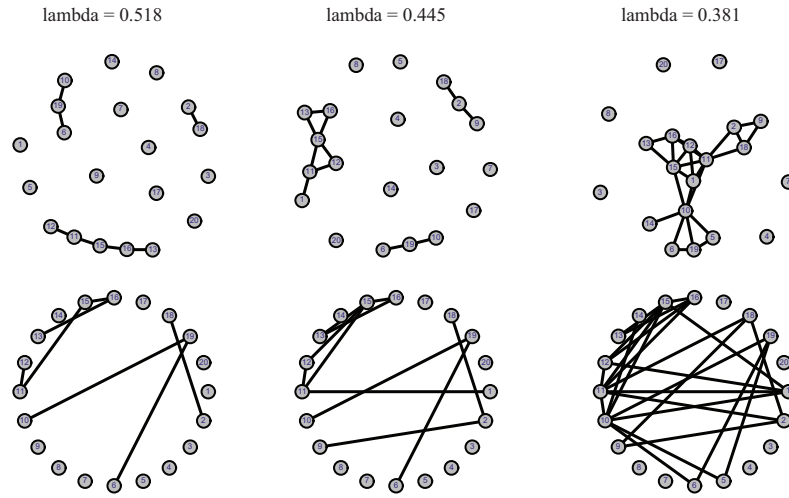


Figure 4: The estimated graph paths using non-paranormal method.

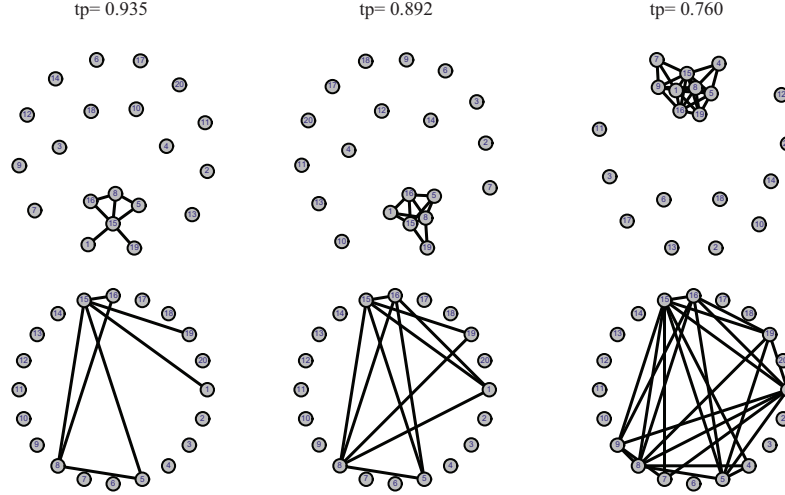


Figure 5: The estimated graph paths using our method. Tuning parameter  $tp$ , is chosen so that the number of edges is close to non-paranormal method that plots figure 4.

Data have been transformed by calculating the log-ratio of the price at time  $t$  to price at time  $t - 1$ , and then standardized by subtracting the mean and adjusting the variance to one.

In order to more distinctly, we sending above output estimated data to R package ggm [3], and plot the graph again, in two different layout, circle and Fruchterman-Reingold. Figure 4 shows the results.

## References

- [Friedman (2007)] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432441, 2007b.
- [Rizzo, and Bakirov (2007)] Szekely, G. and Rizzo, M. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3 (4):12361265.
- [Liu, et al (2009)] Liu, H. , Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* 10 22952328.



- [Zhao, et al (2012)] T. Zhao, H. Liu, K. Roeder, J. Lafferty and L. Wasserman. (2012). The huge Package for High-dimensional Undirected Graph. *Journal of Machine Learning Research* 13, 1059-1062
- [Lin, et al (2009)] Yuanqing Lin, Shenghuo Zhu, Daniel D. Lee, Ben Taskar: Learning Sparse Markov Network Structure via Ensemble-of-Trees Models. *Journal of Machine Learning Research - Proceedings Track 5*: 360-367 (2009)
- [1] J. Fox, et al. (2012). R Package Rcmdr, (Version 1. 8-4)
- [2] M. Rizzo, et al. (2012). R Package energy, (Version 1. 4-0)
- [3] Giovanni M, et al. (2012). R Package ggm, (Version 1. 99-2)

## A Appendix

### Simplified distance correlation

The following R code, calculate simplified distance correlation between two vectors x,y.

```
dcor2<-function(x,y){
  n <- length(x);
  if (n != length(y)) {stop("Sample sizes must be equal")}
  u <- matrix(0,2,n+1);
  w<-0

  for(i in 1:n){
    for(j in 1:n){
      if(i<j){
        w<-abs(x[i]-x[j]); u[1,i]<-u[1,i]+w; u[1,j]<-u[1,j]+w;
      }
      else if(i>j){
        w<-abs(y[i]-y[j]); u[2,i]<-u[2,i]+w; u[2,j]<-u[2,j]+w;
      }
    }
  }
  u <-u/n
  u[1,n+1]<-mean(u[1,][1:n])
  u[2,n+1]<-mean(u[2,][1:n])

  r<-0; rx<-0; ry<-0;
  for(i in 1:n){
    for(j in 1:n){
      r<- r +(abs(x[i]-x[j])-u[1,i]-u[1,j]+u[1,n+1])*(abs(y[i]-y[j])-u[2,i]-u[2,j]+u[2,n+1])
      rx<- rx+(abs(x[i]-x[j])-u[1,i]-u[1,j]+u[1,n+1])*(abs(x[i]-x[j])-u[2,i]-u[2,j]+u[2,n+1])
      ry<- ry+(abs(y[i]-y[j])-u[1,i]-u[1,j]+u[1,n+1])*(abs(y[i]-y[j])-u[2,i]-u[2,j]+u[2,n+1])
    }
  }
  rx<-sqrt(rx)/n
  ry<-sqrt(ry)/n
  r<-sqrt(r)/n
  r/(sqrt(rx*ry))
}
```